智库圆桌（第2期·总211期）

推动人工智能安全发展

人工智能是引领新一轮科技革命和产业变革的重要驱动力。2023年中央经济工作会议提出，要大力推进新型工业化，发展数字经济，加快推动人工智能发展。“十四五”规划和2035年远景目标纲要强调，加强网络安全关键技术研发，加快人工智能安全技术创新，提升网络安全产业综合竞争力。本期特邀专家围绕相关问题进行研讨。

夯实人工智能发展的安全基础

加快推动人工智能发展，需如何应对潜在风险、把握战略主动？

刘鹏

单志广（国家信息中心信息化和产业发展部主任、国家大数据发展专家咨询委员会秘书长）：党的二十大报告提出，推进国家安全体系和能力现代化，坚决维护国家安全和社会稳定。人工智能是引领新一轮科技革命和产业变革的战略性技术，同时也具有明显的“双刃剑”特征。一方面，人工智能赋能网络攻防、开源情报等国家安全相关领域，是筑牢国家安全屏障的有力抓手；另一方面，人工智能因其脆弱性、不稳定性、不可解释性等特点，在与经济社会深度融合应用的过程中，极易引发国家、社会、企业和个人等层面的安全风险。在席卷全球的人工智能浪潮中，如何应对人工智能风险、把握发展战略主动、有效维护和保障国家安全，是国家治理的重要议题。

近年来，国家高度重视人工智能安全发展，逐步完善相关政策法规。国务院印发《新一代人工智能发展规划》提出面向2030年我国新一代人工智能发展的指导思想、战略目标、重点任务和保障措施，部署构筑人工智能发展的先发优势，加快建设创新型国家和世界科技强国。面向算法治理，出台《关于加强互联网信息服务算法综合治理的指导意见》《互联网信息服务算法推荐管理规定》等。面向人工智能合成技术的快速突破，出台《互联网信息服务深度合成管理规定》《生成式人工智能服务管理暂行办法》等。在全球数字经济激烈竞争格局下，科学把握风险防范的尺度至关重要。2023年7月24日中共中央政治局会议强调“促进人工智能安全发展”，体现了坚持统筹发展和安全、坚持发展和安全并重的理念，释放了以人工智能技术激发数实融合新动能、打造高质量发展新引擎的积极信号。

推动人工智能产业快速发展，要把保障数据安全放在突出位置。

第一，数据是数字经济时代的关键生产要素，保障数据安全是促进人工智能安全发展的重要基础。我国2022年数字经济规模逾50万亿元，总量稳居世界第二，占GDP比重41.5%，数据量呈爆发式增长态势。随着数据要素规模不断扩大，以人工智能为代表的数字技术将实现知识与数据双轮驱动，数据价值得到进一步释放，生产资源配置、生产运营逻辑以及生产、分配、流通和消费关系等得以重塑，生产方式和生产关系发生变革，赋能传统产业转型升级，助力数字经济快速发展。同时，也伴随着数据泄露、虚假信息、算法歧视等数据安全新问题。只有筑牢数字安全屏障，才能为人工智能发展保驾护航。

第二，人工智能产业快速发展过程中显现出数据安全领域的风险挑战。当前，人工智能进入快速发展期，应高度关注并有效应对随之而来的问题。例如，神经网络具有“黑盒”特点，导致人工智能存在不可解释性；深度学习对训练样本过度依赖，导致学习结果的不可判定性；神经网络前向推进的不可逆，导致结果的不可推论性。此外，漏洞、后门等引发的问题交织叠加，使得人工智能应用系统的数据安全问题变得更加复杂。针对IT行业领导者进行的一项关于Chat-GPT等大模型的调查显示，安全性是受访者最关心的问题，71%的受访者认为生成式人工智能会给企业的数据安全带来新的风险。为了防止敏感数据外流，微软、亚马逊等科技公司已相继限制或禁止其员工使用生成式人工智能工具。可见，全面加强人工智能数据安全保障体系和能力建设已成为应对新形势新挑战的必然之举。

人工智能时代的数字安全威胁到底有多大？一方面，人工智能系统自身面临多维度安全风险。技术内生风险和系统衍生风险交织叠加，使得人工智能时代的安全问题异常复杂。数据安全风险方面，人工智能依托海量数据发展，有敏感信息泄露风险，且人工智能平台收集的原始数据与衍生数据的归属权、控制权和使用权目前在法律上尚难界定；算法模型安全方面，安全风险贯穿数据采集、预处理、模型训练、模型微调、模型部署应用等人工智能模型构建的全生命周期；外部攻击安全方面，数据投毒、模型后门、对抗样本、数据泄露、模型窃取、软件漏洞等安全隐患屡见不鲜。

另一方面，人工智能技术滥用带来数字安全威胁。当前，生成式人工智能的发展标志着人工智能正在从专用智能迈向通用智能，进入了全新发展阶段。大部分传统人工智能模型的安全风险仍然存在，同时生成式人工智能也有一些特有的问题：技术软肋难以避免，易培育假信息“温床”；使用方式简单便捷，易形成失泄密“陷阱”；新兴技术尚难监管，易成为信息战“武器”。

因此，亟需加强人工智能发展的潜在风险研判和防范，确保人工智能安全、可靠、可控。

完善数据安全监管体系

我国在创新人工智能技术手段及完善数据安全监管方面取得哪些成效？

陈凤仙（中国电信研究院高级分析师）：党中央高度重视人工智能安全发展问题，围绕产业发展、科技伦理、算法治理及行业应用安全等方面，加快人工智能安全技术创新，逐步形成一套较为完善的发展政策和法规体系，确保维护国家安全和社会公共利益。据统计，我国人工智能核心产业规模达5000亿元，算力总规模位居全球第二。

在规范人工智能科技伦理方面，陆续发布《新一代人工智能治理原则——发展负责任的人工智能》《新一代人工智能伦理规范》等，积极引导全社会负责任地开展人工智能研发和应用。2023年5月，工信部科技伦理委员会、工信领域科技伦理专家委员会正式成立，进一步加强科技伦理审查和监管。

在强化人工智能算法治理方面，我国在规制生成式人工智能领域率先推出多项有力举措。国家互联网信息办公室等部门2022年11月联合发布《互联网信息服务深度合成管理规定》，明确生成式人工智能应用服务提供者、技术支持者和服务使用者等各方法定义务；2023年7月联合发布《生成式人工智能服务管理暂行办法》，鼓励生成式人工智能创新发展，对生成式人工智能服务实行包容审慎和分类分级监管。

在人工智能全球治理合作方面，我国积极参与、多方实践，取得重要进展。2022年11月，我国向联合国《特定常规武器公约》缔约国大会提交《中国关于加强人工智能伦理治理的立场文件》，提出人工智能治理要坚持伦理先行、加强自我约束、强化责任担当、鼓励国际合作等多项主张，表明了推动各方共商共建共享、加强全球治理、积极构建人类命运共同体的中国立场。2023年4月，我国向联合国提交《中国关于全球数字治理有关问题的立场》，明确表示各国应在普遍参与的基础上，通过对话与合作，推动形成具有广泛共识的人工智能国际治理框架和标准规范。

与此同时，我国将数据安全放在保障人工智能安全发展的突出位置，大力推动数据资源建设，强化安全防护技术手段迭代升级，完善数据安全监管体系，取得明显成效。

稳步推进数据基础制度构建、数据资源供给和流通利用。在国家层面，逐步完善数据资源建设顶层设计，对加快培育统一的技术和数据市场及经营主体、构建数据基础制度等提出明确要求。2023年10月，国家数据局挂牌成立，推动数据实现从自然资源到经济资产的跨越。在地方层面，多地加快培育规范数据交易市场。2022年1月，北京国际大数据交易所率先在全国建立数字经济中介产业体系。截至2023年9月，全国注册成立的数据交易机构已有60家。2023年以来，北京、上海、广东、江西、湖北、贵州等密集发布政策文件，深化数据要素市场改革和创新。为助力大模型应用落地，一些地方积极探索政产学研联合构建高质量数据集。例如，北京启动实施通用人工智能产业创新伙伴计划，发布“北京市人工智能大模型高质量数据集”。国家互联网信息办公室发布《数字中国发展报告（2022年）》显示，2022年我国数据产量达8.1ZB，同比增长22.7%，全球占比10.5%，位居世界第二。

大幅提升依靠技术解决安全风险问题的能力。一方面，网络安全技术创新应用活跃。近年来，相关机构持续强化网络安全技术布局及应用，加快推动重点领域和细分环节技术突破、专利布局和标准转化。另一方面，网络安全、数据安全产业快速发展。2022年，我国网络安全产业规模增速约为13.9%。北京、长沙、成渝三大国家网络安全产业园区相继成立，汇聚网络安全企业超500家，10个网络安全创新应用示范区加速建设。2023年1月，工信部等部门印发《关于促进数据安全产业发展的指导意见》，提出到2025年数据安全产业规模超1500亿元，年复合增长率超30%，推动数据安全产业驶入快车道。

加快推进立法进程，标准体系建设取得阶段性成效。一方面，完善数据安全监管法律依据，推动重要数据和个人信息保护合规水平进一步提升。建立应对数据泄露等事故的应急响应机制，及时启动应急预案并妥善处置。另一方面，抓紧研制数据质量、数据安全、算法正确性等技术规范和标准。2023年8月，我国发布人工智能安全基础标准《信息安全技术 机器学习算法安全评估规范》。同时，在生物特征识别、智能汽车等人工智能关键应用领域发布多项国家标准，支撑人工智能安全发展。

强化生成式人工智能安全防范

生成式人工智能引发新一轮智能化浪潮，筑牢数字安全屏障需采取哪些新举措？

武虹（中国科协创新战略研究院研究员）：生成式人工智能作为大模型、大数据、大算力的产物，在强大算力支持下，借助大型语言模型将收集到的海量信息进行处理并回应用户的个性化需求，近乎无成本地生成针对特定提问内容编写的答案，是人工智能领域一次出色的集成创新。

也应看到，生成式人工智能并不公开其收集和整理的海量互联网语料、模型架构及训练内容，加之其深度学习算法基于统计原理之上，仅对客观规律进行揭示却无法给予解释，已构成事实上的数据及技术黑箱，势必会对数据安全模式带来新的扰动。因此，有必要采取新的应对措施。

一是强化数据全生命周期监督管理。生成式人工智能语料库的大规模集聚将带来包含数据采集、处理加工、存储及输出等覆盖数据全生命周期的安全问题。数据采集输入阶段，可能会有未经审核的虚假信息入库进而影响语料库质量；数据处理加工阶段，语料库的标注过程可能有意识形态和价值观影响风险，同时算法也可能会有相当程度的倾向性引导，其运算结果又会被潜移默化地注入后续数据处理；数据输出阶段，更是存在非真实世界批量自动产生的海量数据被当作新的语料库，产生后续迭代风险。因此，建议进行数据全生命周期的安全体系构建。例如，针对个人隐私及知识产权等数据，通过隐私计算及区块链等强化数据安全防范力度；针对医疗、金融、电商等重点行业，通过访问控制、安全可信计算环境等技术手段加强防护。

二是提升对攻击性人工智能的防范意识。攻击性人工智能通常分为两种形式，即“使用人工智能的攻击”和“攻击人工智能”。随着生成式人工智能技术快速提升，网络攻击者编写恶意代码以及实施数据攻击的技术门槛大大降低。同时，大模型也面临被注入特定引导词以诱导其输出伪造数据甚至违法答案等间接数据安全风险。传统的成本高昂的攻击手法，向分布式、智能化、自动化方向演进。建议推动以企业为主体、基于人工智能的新一代数据安全防护等专项研究；鼓励相关行业的科技领军企业发布横向课题，联合高校及科研院所开展协同攻关；通过风险投资引导初创公司将成果应用于数据安全对抗业务，促进以企业应用为导向的生成式人工智能对抗模型产学研一体化创新体系构建。

三是完善国家层面的数据安全战略规划及顶层设计。生成式人工智能在进行人类对话、推理和翻译写作时，给人类的信息掌控及自主决策能力等带来挑战。同时，生成式人工智能也极大促进了交互式数据的迭代输出与自动传输，增加了危及国家数据主权、信息与网络空间安全的潜在风险。面对生成式人工智能引发的不确定性，需提前研判可能的安全风险，建立健全政府、企业与社会等沟通交流机制，探索推动多方合作的治理模式，加速构建国家层面数据安全战略规划和大模型监管应用法律支撑，通过夯实自主可控新基建设施、加快行业自治规范与国家强制性法律法规等协同体系构建，以及发起或加入单边及多边协议或联盟等方式，巩固国家数据安全防线。

从国际上看，在生成式人工智能安全防范方面，一些国家的经验做法值得借鉴。例如，欧盟2021年提出《人工智能法案》草案，旨在基于风险识别分析方法为人工智能制定统一的法律监管框架和规制体系。2023年12月，欧洲议会、欧盟委员会和27个成员国谈判代表就该法案达成协议，针对ChatGPT等生成式人工智能工具的透明度问题做出相应规定，其中包括由人工智能生成的内容需提供受版权保护的训练数据集摘要等，还对人工智能风险级别进行了划分，并给出对应的监管要求。美国白宫2022年发布《人工智能权利法案蓝图》，将公平和隐私保护视为法案的核心宗旨。2023年1月，美国国家标准技术研究院发布人工智能风险管理框架，提供系统化评估路径，将人工智能的风险管理分为治理、映射、测量和管理4个模块。其中，治理模块主要针对人工智能系统全生命周期实行有效风险管理机制；映射模块主要用于明确特定场景与其对应的人工智能风险解决方案；测量模块主要采用定量、定性或混合工具，对人工智能系统风险和潜在影响进行分析、评估、测试和控制；管理模块主要针对系统风险进行判定、排序和响应，明确风险响应步骤，定期监控记录并完善风险响应和恢复机制。

可见，生成式人工智能更要兼顾发展与安全，重视防范风险与包容审慎平衡，从而更好推动经济高质量发展。

探索

以ChatGPT为代表的生成式人工智能掀起新一轮热潮。与此同时，数据泄露、隐私窃取、算法歧视等数字安全风险不断显现，迫切需要寻找共享与监管并重的动态平衡范式，守住人工智能时代的数字安全底线。

纵观全球，中国、美国和欧盟作为探索数字安全和数字治理的先行者，无论是技术创新还是立法规范都走在世界前列，同时也存在差异。在相同点方面，均高度重视算法治理，将算法安全嵌套在数据安全中，实行数据与算法协同治理；在差异性方面，虽然同样强调个人隐私安全，美国以鼓励创新为核心，更注重数据自由流动，倾向于以行业自律进行治理。欧盟注重个人隐私保护和立法，探索和引入人工智能监管沙盒机制，并发布首部人工智能监管法案。这些经验做法，对我国数字安全治理具有一定参考价值。

我国加快推动人工智能发展，需形成政府、企业、社会组织和个人合力，在协同数据和算法治理、保障生成式人工智能安全等方面实现重点突破。

第一，启动国家人工智能数据和算法工程。建立安全标准，分门别类对数据和算法进行管理，提升数据互操作性以及算法透明度，改变过去个人或企业单打独斗的局面。

自动检索风险指标。开发针对人工智能应用网络的早期预警系统，对网络资源进行常规监控和过滤，对于不符合政策要求的危险因素、劣质数据和不良信息，及时清除或屏蔽。通过该预警系统自动检索关键风险指标，及时补救，防止再出现类似安全漏洞。

甄别人工智能生成内容。从源头上，为人工智能生成内容打上标记。深度合成服务提供者提供深度合成服务，可能导致公众混淆或者误认的，应在生成或编辑的信息内容的合理位置、区域进行显著标识，向公众提示深度合成情况。

加大人工智能生成内容检测工具开发和优化。目前针对人工智能生成的图像、文本等已出现相应检测工具，用于区分人工智能生成的内容和人类创造的内容，但准确率不高。亟需加大对数据、算法、模型的研究，开发精准的生成式人工智能检测工具，真正实现“以AI测AI”。

加强人工智能数据共享。数据、算法、算力是驱动人工智能发展的“三驾马车”，其中数据是人工智能发展的养料，例如仅GPT-4的训练数据集就包含约13万亿个词元。如果缺乏足够的数据，人工智能发展无异于“无米之炊”。

推动中文数据集共享。由于语言特点、获取成本、开源程度以及数据集质量要求等原因，相较于英文数据，目前中文数据集规模较小。基于中文的人工智能开发，可通过国家人工智能数据工程汇总高质量中文数据集，并促进数据分类分级有序共享，使安全性和服务质量得到大幅提升。

第二，加强生成式人工智能监管。推进全球沟通和探讨，通过网络数据安全管理、个人信息保护、数据审计等法律法规进一步完善生成式人工智能监管。统筹数据安全与算法治理，针对金融、医疗等不同行业领域以及算法歧视、算法黑箱等问题，开展多层次和精细化监管。开展多模态智能分析，在大模型领域引入文本、图像、语音等，在训练和应用过程中细化对不同元素的监管，通过功能模块设计，及时发现问题并防范风险。

切实保障数据安全。使用生成式人工智能产品过程中，也在同步收集用户数据和信息，可能引发潜在隐私安全问题。对此，应扩大安全使用指南宣传。例如，不主动分享敏感信息、关闭聊天记录等，基于专门的云服务运行，从访问控制、数据加密、网络连接等方面加强保护。对于数据敏感度较高的用户，通过敏感信息过滤一体机进行识别筛选，可有效避免大模型产品在提供服务时产生不可控信息。

构建特定知识库。基于特定知识库提供人工智能服务，可在一定程度上避免虚假错误信息，提升准确性和安全性。建议借助大模型训练推理一体机，通过本地化训练和推理，在保护用户数据隐私的前提下构建特定知识库。

第三，在国际层面，积极与联合国及主要国家沟通交流，达成全球规避人工智能风险共识，推动对所有大型人工智能科研项目实施备案和风险评估制度。在国家层面，组织专家团队潜心进行人工智能风险评估和研究相关立法，加强网络安全关键技术研发，加快人工智能安全技术创新，提升网络安全产业综合竞争力。

（作者系中国大数据应用联盟人工智能专家委员会主任）

本版编辑 张静美 编 王子萱
来稿邮箱 jjrbjjzk@163.com